"Express Mail" mailing label number EV 331533802 US

Date of Deposit: October 28, 2003

Attorney Docket No.: 15153US01

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

CLASSIFICATION OF SPEECH AND MUSIC USING ZERO CROSSING

FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0001] [Not Applicable]

[MICROFICHE/COPYRIGHT REFERENCE]

[0002] [Not Applicable]

BACKGROUND OF THE INVENTION

[0003] Human beings, with normal hearing, are often able to distinguish sounds from about 20 Hz, such as the lowest note on a large pipe organ, to 20,000 Hz, such as the high shrill of a dog whistle. Human speech, on the other hand, ranges from 300 Hz to 4,000 Hz.

[0004] Music may be produced by playing musical instruments. Musical instruments often produce sounds that lie outside the range of human speech, and in many instances, produce sounds (overtones, etc.) which lie outside the range of human hearing.

[0005] An audio communication can comprise either music, speech or both. However, conventional equipment processes audio communication signals comprising only speech in a similar manner as communication signals comprising music.

[0006] Further limitations and disadvantages of conventional and traditional approaches will become apparent to one of skill in the art, through comparison of such systems with embodiments presented in the remainder of the present application with references to the drawings.

SUMMARY OF THE INVENTION

[0007] Aspects of the present invention may be found in a method for classifying an audio signal. The method may comprise receiving an audio signal to be classified, analyzing selected audio signal components, recording a result of analysis of the selected audio signal components, comparing the recorded result of analysis to a threshold value, and classifying the audio signal based upon comparison of the recorded result of analysis and the threshold value.

[0008] In another embodiment of the present invention, classifying the audio signal based upon comparison of the recorded result of analysis and the threshold value may further comprise: if the recorded result of analysis is greater than the threshold value, then the audio signal is determined to be music; and if the recorded result of analysis is less than the threshold value, then the audio signal is determined to be speech.

[0009] In another embodiment of the present invention, analyzing the selected audio signal components may comprise counting zero point transitions of the selected audio signal components.

[0010] In another embodiment of the present invention, recording a result of analysis of the selected audio signal components may comprise recording a count value of a number of zero point transitions of the selected audio signal components.

[0011] In another embodiment of the present invention, transmitting components of the audio signal having a frequency less than a predetermined frequency may comprise passing the audio signal through a low pass filter. The low pass filter may be adapted to permit transmission of frequencies below the predetermined frequency.

[0012] In another embodiment of the present invention, selecting a number of transmitted audio signal components for analysis comprises passing transmitting digital audio components through a decimator. Every 1 in N audio signal components may be transmitted and audio signal components between 1 and N may be discarded.

[0013] In another embodiment of the present invention, classifying the audio signal may further comprise turning on a flag in a header of a packet of digital audio information. The flag provides an indication of classification of the audio signal based upon comparison of the recorded result of analysis and the threshold value.

[0014] In another embodiment of the present invention, the method may further comprise transmitting components of the audio signal having a frequency less than a predetermined frequency and selecting a number of transmitted audio signal components for analysis.

[0015] In another embodiment of the present invention, classifying the audio signal may occur at a transmitting end of an audio transmission system.

[0016] In another embodiment of the present invention, classifying the audio signal may occur at a receiving end of an audio transmission system.

[0017] In another embodiment of the present invention, the audio signal is one of an analog signal and a digital signal.

[0018] In another embodiment of the present invention, the threshold value used in the comparison is pre-determined and pre-set by a user.

[0019] In another embodiment of the present invention, the threshold value used in the comparison determined through trial and error of a plurality of iterations in a comparing device.

[0020] In another embodiment of the present invention, analyzing selected audio signal components may comprise counting zero point transitions of the audio signal for a predetermined period of time.

[0021] In another embodiment of the present invention, the method may further comprise converting the audio signal from an analog signal to a digital signal, encoding the audio signal, packetizing the audio signal, transmitting the audio signal, decoding the audio signal, and processing the audio signal. Processing may at least comprise one of storing the audio signal and playing the audio signal.

[0022] Aspects of the present invention may also be found in an apparatus for classifying an audio signal. The apparatus may comprise a zero point counter for counting and recording zero point transitions encountered in analysis of the selected audio signal components and a comparator for comparing a recorded result of analysis to a threshold value and classifying the audio signal based upon comparison of the recorded result of analysis and the threshold value.

[0023] In another embodiment of the present invention, classifying the audio signal based upon comparison of the recorded result of analysis and the threshold value in the comparator may further comprise: if the recorded result of analysis is greater than the threshold value, then the audio signal is determined to be music; and if the recorded result of analysis is less than the threshold value, then the audio signal is determined to be speech.

[0024] In another embodiment of the present invention, the apparatus may further comprise a low pass filter for preventing transmission of components of the audio signal having a frequency greater than a predetermined frequency and a decimator for selecting a reduced number of audio components for analysis.

[0025] In another embodiment of the present invention, the decimator selecting a reduced number of audio components for analysis may further comprise the decimator selecting every 1 in N audio signal components to be transmitted and selecting the audio signal components between 1 and N to be discarded.

[0026] In another embodiment of the present invention, the apparatus may further comprise at least one of an audio signal encoder and an audio signal decoder.

[0027] In another embodiment of the present invention, the apparatus may further comprise a speech/music classifying device being associated with the audio signal encoder.

[0028] In another embodiment of the present invention, the apparatus may further comprise a speech/music classifying device associated with the audio signal decoder.

[0029] In another embodiment of the present invention, the apparatus may further comprise a signal processor and an audio processing unit associated with the audio signal decoder.

[0030] In another embodiment of the present invention, the apparatus may further comprise a bitstream multiplexer associated with the audio signal decoder.

[0031] These and other advantages and novel features of the present invention, as well as details of an illustrated embodiment thereof, will be more fully understood from the following description and drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0032] **Figure 1** illustrates a portion of an audio communication received by an electronic device according to an embodiment of the present invention;

[0033] **Figure 2** illustrates a portion of an analog audio signal according to an embodiment of the present invention;

[0034] **Figure 3** illustrates a portion of an analog audio signal being sampled for conversion to a digital signal according to an embodiment of the present invention;

[0035] **Figure 4** illustrates a portion of a digital audio signal according to an embodiment of the present invention;

[0036] **Figure 4A** is a flowchart illustrating a method of classifying whether an audio communication is speech or music according to an embodiment of the present invention;

[0037] **Figure 5** illustrates an apparatus for classifying an audio signal as either speech or music using zero crossing analysis according to an embodiment of the invention;

[0038] **Figure 6** is a flow chart illustrating an exemplary processing method performed by the apparatus of **Figure 5** for classifying an audio signal as speech or music using a zero crossing counting method according to an embodiment of the present invention;

[0039] **Figure 7** is a block diagram illustrating a system for converting, classifying, encoding, and packetizing an audio communication according to an embodiment of the present invention;

[0040] **Figure 8** is a block diagram illustrating encoding of an exemplary audio signal A(t) according to an embodiment of the present invention; and

[0041] **Figure 9** is a block diagram illustrating an exemplary audio decoder according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[0042] Modern electronic devices are adapted to transmitting and receiving both music and speech. In audio communication, any interruption of music transmission, such by speech transmission, may be interpreted as a commercial or an advertisement, or vice versa.

[0043] An aspect of the present invention may be found in a method and system for classifying whether a communication received is speech or music by applying a zero crossing analysis method to the communication.

[0044] **Figure 1** illustrates a portion 100 of an audio communication 110 received by an electronic device according to an embodiment of the present invention. The audio communication 110 comprises an analog or digital audio signal having a bandwidth or spectrum. The audio communication 110 oscillates between positive amplitude maxima 101 and negative amplitude maxima 103, crossing a zero point 109 (zero point crossings 105 marked by X's) as each oscillation transitions from positive to negative values. The audio communication 110 is illustrated in terms of the amplitude 108 (Y-Axis) with respect to time 106 (X-axis).

[0045] **Figure 2** illustrates a portion 200 of an analog audio signal 210. The analog audio signal 210 comprises a bandwidth or spectrum. The analog audio signal 210 oscillates between a positive amplitude 201 and a negative amplitude 203, crossing a zero point 209 (the zero point crossing 205 marked by an X) as each oscillation transitions from positive to negative values. The analog audio signal 210 is illustrated in terms of the amplitude 208 (Y-Axis) with respect to time 206 (X-axis).

[0046] **Figure 3** illustrates a portion 300 of an analog audio signal 310 being sampled for conversion to a digital signal according to an embodiment of the present invention. The audio signal 310 comprises a bandwidth or spectrum and has been divided into a plurality of discrete samples 312. The samples 312 approximate the analog audio signal 310. The analog audio signal 310 oscillates between a positive amplitude 301 and a negative amplitude 303, crossing a zero point 309 (the zero point crossing 305 marked by an X) as each oscillation transitions from positive to negative values. The sampled audio signal 310 is illustrated in terms of the amplitude 308 (Y-Axis) with respect to time 306 (X-axis).

[0047] **Figure 4** illustrates a portion 400 of a digital audio signal 410 according to an embodiment of the present invention. The digital audio signal 410 comprises a bandwidth or spectrum and is shown approximating the analog signal 210 through a plurality of quantized discrete samples 412. The digital audio signal 410 transitions through a positive amplitude 401 and a negative amplitude 403 over time, crossing a zero point 409 (the zero point crossing 405 marked by an X). The digital audio signal 410 is illustrated in terms of the quantized amplitude 408 (Y-Axis) with respect quantized time 406 (X-axis).

[0048] A digital audio signal is an audio signal using binary code to represent audio information. The signals are modeled so that the information being transmitted is translated into a series of zeros and ones, i.e., a range of analog values are associated with a logical value. Digital systems process time varying signals that can take on any value quantized from a continuous range of electrical values. The digital audio transmission system takes the audio information and represents it as a series of bits represented in code by zeros and ones.

[0049] On the other hand, an analog audio communication is a way of sending signals in which the communicated audio signal is a wave reflecting the original signal. An analog audio communication system attempts to recreate the audio information as it actually happens. Analog systems process time varying signals that can take any value across a continuous electrical values.

[0050] Human beings with normal hearing can detect sounds from about 20 Hz to about 20,000 Hz. Human speech, on the other hand, ordinarily ranges from about 300 Hz to about 4,000 Hz. Music produces audible sounds that lie outside the range of human speech (20 to 20,000 Hz) but within the range of human speech (300 to 4,000 Hz).

[0051] There are various reasons for determining whether the audio communication is associated with speech or music. For example, it may be advantageous to process audio communications associated with speech in one manner and audio communications associated with music in another manner.

[0052] Whether the audio communication is associated with speech or music can be determined by measuring the number of times the audio signal crosses the zero point (zero point crossing) during a given period of time. The higher the number of zero point crossings 105, the greater the likelihood that the audio communication is associated with music, while the lower the number of zero point crossings 105, the greater the likelihood that the audio communication is associated with speech.

[0053] Accordingly, the number of zero point crossings can be compared to a threshold. If the number of zero point crossings exceeds a predetermined threshold value which can be computed offline by analyzing the given audio signal, a determination can be made that the audio communication is associated with music. If the threshold value exceeds the number of zero point crossings, a determination is made tat the audio communication is associated with speech.

[0054] **Figure 4A** is a flowchart 400A illustrating a method of classifying whether an audio communication is speech or music according to an embodiment of the present invention. At block 410A, the flowchart illustrates measuring the number of zero crossings during a given period of time. At block 420A, the flowchart illustrates comparing the number of zero crossings to a threshold value. At decision block 430A, the result of the comparison is determined and the question of whether the number of zero crossings exceeds the threshold value is answered. If the number of zero crossings is greater than the threshold value (Yes), then the audio signal is determined to be music 440A. However, if the number of zero crossings is less than the threshold value (No), then the audio signal is determined to be speech 450A.

[0055] **Figure 5** illustrates an apparatus 500 for classifying an audio signal as either speech or music using zero crossing analysis according to an embodiment of the invention. The apparatus 500 comprises an input 520, a low pass filter 530, a decimator 540, a zero point counter 550, a comparator 560, and an output 570. An exemplary signal processing method performed by the apparatus will be described in detail in **Figure 6**.

[0056] **Figure 6** is a flow chart 600 illustrating an exemplary processing method performed by the apparatus of **Figure 5** for classifying an audio signal as speech or music using a zero crossing counting method according to an embodiment of the present invention. In order to classify the audio signal illustrated in **Figure 1** as speech or music, the audio signal may be passed through a low pass filter 610. The low pass filter may be a filter, which permits transmission of audio signals having a frequency between 0 and 4,000 Hz, while blocking or preventing those audio signals having a frequency greater than 4,000 Hz from being transmitted.

[0057] The low pass filter 530 permits analysis of audio that may be characteristic of human speech because that portion of the audio signal spectrum outside the range of human speech has been filtered from further transmission by the low pass filter 530. Thus, the low pass filter 530 also reduces the amount of audio information to be analyzed by limiting the information to that which may at least comprise human speech.

[0058] The filtered signal, if digital, may also be passed (620) through a decimator 540. The decimator 540 further limits the amount of audio information to be analyzed by reducing the resolution of the digital audio signal. The decimator may be adapted to permit transmission of one audio signal transition (i.e., sample) in N, where N may be an integer selected to provide a particular level of discrimination.

[0059] The portions of the audio signal not selected for further analysis, i.e., those audio signal transitions between 1 and N, may be discarded. After passing the signal through the decimator 540, the amount of audio signal information to be analyzed has been further reduced.

[0060] The audio signal information may be passed (630) through a zero point counter 550. In the zero point counter 550, every time the audio signal transitions from positive to negative value or from negative to positive value, the audio signal crosses the zero point boundary, a count is advanced (640) one integer count. When an audio signal over a predetermined time interval has been zero point counted, or when the counting has taken place for a predetermined amount of time, the recorded count value is transmitted (650) to a comparator 560.

[0061] In the comparator 560, the recorded count value is compared (660) to a threshold count value 660. The comparator determines if the recorded count is greater than the threshold value 666. If the recorded count value is greater than the threshold count value (Yes), then the audio signal is determined to be music 670, however, if the recorded count value is less than the threshold count value then (No), the audio signal is determined to be speech 680.

[0062] The comparator 560 may comprise at least one buffer for storing audio signal information during comparison. The comparator 560 may be adapted to process the signal with even finer discrimination, i.e., determine more about the signal than just whether the signal is music or speech. For example, if the signal is determined to be speech, the frequency range compatible with human speech may be further compared to a sub-threshold value to determine if the speech is male speech, female speech, adult speech, or child speech based upon the number of zero crossings the signal comprises in a particular corresponding frequency range.

[0063] Additionally, if the signal is determined to be music, a different sub-threshold value may be used to determine what characteristic instrument(s) are making the music based upon the zero crossings the signal comprises in a particular corresponding frequency range.

[0064] In general, the dominant classifying sub-band, as determined from the comparison of the number of zero crossings to the threshold value, may be further divided and mathematically analyzed to glean additional information about the identity of the producer of the sound represented by the audio signal.

[0065] The threshold value may be predetermined and provided by a user, or alternatively may be learned through a training process in the comparator, wherein the comparator, through trial and error, determines the threshold value. The comparator may compare the zero crossing count to the threshold value and output a classification of the audio signal as being one of music or speech.

[0066] An audio signal comprising human speech has fewer zero point crossings than one comprising music, and thus a lower recorded count value. **The reason the reason the audio signal comprising human speech has fewer zeros crossings is a result of the physical size of the human vocal tract, which is unable to oscillate beyond a certain frequency. The human vocal tract produces sound having a limited fundamental frequency (i.e., pitch). Speech harmonics are mostly restricted to below 4 KHz, i.e., most of the speech audio signal energy lies within a 0 to 4 KHz spectrum.**

[0067] **Figure 7** is a block diagram illustrating a system 700 for converting, classifying, encoding, and packetizing an audio communication according to an embodiment of the present invention. In **Figure 7**, the system 700 receives an audio communication 710, wherein the audio communication may be either an analog signal 701 or a digital signal 703. The audio signal 710 may proceed directly to speech/music classification apparatus 766 as an analog signal 701 at junction 763. Alternatively, the audio signal 710 may be passed through analog to digital converter 705 for conversion to a digital signal 703 that is provided via junction 797 to the speech/music classification apparatus 766. After conversion from analog to digital, the digital signal 703 may be passed to MPEG encoder 725. The circumstances of the audio signal processing at the MPEG encoder will be described below.

[0068] The audio signal may arrive at the speech/music classifying apparatus 766 at input 720. The signal is then passed through low pass filter 730 where those frequencies above 4,000 KHz (i.e., those frequencies outside the range of human speech) are discarded. If the signal is an analog signal 701, decimator 740 is by-passed and the signal is passed directly from the low pass filter 730 to the zero point counter 750. However, if the signal is a digital signal 703, the signal is passed to the decimator 740 and the amount of data is further reduced. Only a digital signal, may be processed by decimator 740. At the decimator 740, 1 in N samples are retained, while all the intervening samples are discarded. N may be chosen to be any desired integer and may be determined in advance by a user.

[0069] When the signal arrives at the zero point counter 750, the zero point transitions (each time the signal crosses the zero point) are counted. The zero point counter 750 continues to count zero crossings for a predetermined period of time. After the predetermined period of time has expired, a zero crossing count value is passed to comparator 760. Comparator 760 is adapted to compare the zero crossing count value to a threshold value. The threshold value may be pre-set by a user, or the comparator may determine (learn) the threshold value through trial and error. If the zero crossing count value is greater than the threshold value, then the output from the speech/music classifying apparatus 766 is that the audio signal is determined to be music. However, if the zero crossing count value is less than the threshold value, then the output from the classifying apparatus 766 is that the audio signal is speech.

[0070] The signal may then be passed to either MPEG encoder 725 or alternatively to packetization engine 735 via junction 795. The MPEG encoder 725 converts the digital signal 703 to an audio elementary stream (AES) encoding the digital signal in accordance with the MPEG standard. When the AES is directed to the packetization engine 735, the AES is packetized into a packetized audio elementary stream comprising packets 755. Each packet comprises a portion of the AES and may also comprise a flag 775. The flag 775 may indicate that the portion of the AES in the packet is speech or music depending upon the state of the flag, i.e., whether the flag is turned on or off.

[0071] **Figure 8** is a block diagram 800 illustrating encoding of an exemplary audio signal A(t) 810 by the MPEG encoder 725 according to an embodiment of the present invention. The audio signal 810 is sampled and the samples are grouped into frames 820 ($F_0 \ldots F_n$) of 1024 samples, e.g., ($F_x(0) \ldots F_x(1023)$). The frames 820 ($F_0 \ldots F_n$) are grouped into windows 830 ($W_0 \ldots W_{n)}$ that comprise 2048 samples or two frames, e.g., ($W_x(0) \ldots W_x(2047)$). However, each window 830 W_x has a 50% overlap with the previous window 830 W_{x-1}.

[0072] Accordingly, the first 1024 samples of a window 830 W_x are the same as the last 1024 samples of the previous window 830 W_{x-1}. A window function w(t) is applied to each window 830 ($W_0 \ldots W_n$), resulting in sets ($wW_0 \ldots wW_n$) of 2048 windowed samples 840, e.g., ($wW_x(0) \ldots wW_x(2047)$). The modified discrete cosine transformation (MDCT) may be applied to each set ($wW_0 \ldots wW_n$) of windowed samples 840 ($wW_x(0) \ldots wW_x(2047)$), resulting sets ($MDCT_0 \ldots MDCT_n$) of 1024 transformation frequency coefficients 850, e.g., ($MDCT_x(0) \ldots MDCT_x(1023)$). **Although an MDCT transformation has been described for purposes of example, other mathematical transformations may be used as processing requires. For example, Fast Fourier Transformation (FFT), Wavelet transformation, etc., may be used to compute the frequency components for the audio signal rather than restricting computation to MDCT transform coefficients. Transformation coefficients may be referred to as coefficients $T_0 \ldots T_N$.**

[0073] The MPEG encoder receives the output of the speech/music classification apparatus. Based upon the output of the speech/music classification apparatus, the MPEG encoder 725 can take any number of actions with respect to the transformation coefficients $\mathbf{T_0 \ldots T_N}$. For example, where the output indicates that the content associated with the audio signal 810 is speech, the MPEG encoder 725 can either discard or quantize with fewer bits the transformation coefficients $\mathbf{T_0 \ldots T_N}$ associated with frequencies outside the range of human speech, i.e., exceeding 4 KHz. Where the output indicates that the content associated with the audio signal 810 is music, the MPEG encoder 775 can quantize the transformation coefficients **$T_0...T_N$** associated with frequencies outside the range of human speech.

[0074] The sets of transformation coefficients **$T_0...T_N$** may then be quantized and coded for transmission, forming what is known as an audio elementary stream (AES). The AES can be multiplexed with other AESs. The multiplexed signal, known as the Audio Transport Stream (Audio TS) can then be stored and/or transported for playback on a playback device. The playback device can either be local or remotely located.

[0075] Where the playback device is remotely located, the multiplexed signal is transported over a communication medium, such as the Internet. During playback, the Audio TS is de-multiplexed, resulting in the constituent AES signals. The constituent AES signals are then decoded, resulting in the audio signal.

[0076] Alternatively, the transformation coefficients **$T_0...T_N$** may be packetized by the packetization engine of **Figure 7**. In an audio signal, each frame may comprise transformation coefficients **$T_0...T_N$**. Sub-frame contents may correspond to a particular range of audio frequencies.

[0077] **Figure 9** is a block diagram illustrating an exemplary audio decoder according to an embodiment of the present invention. Referring now to **Figure 9**, once the frame synchronization is found and delivered from signal processor 901, the advanced audio coding (AAC) bitstream 903 is de-multiplexed by a bitstream de-multiplexer 905. This includes Huffman decoding 916, scale factor decoding

915, and decoding of side information used in tools such as mono/stereo 920, intensity stereo 925, TNS 930, and the filterbank 935.

[0078] The sets of transformation coefficients $\mathbf{T_0...T_N}$ are decoded and copied to an output buffer in a sample fashion. After Huffman decoding 916, an inverse quantizer 940 inverse quantizes each set of transformation coefficients $\mathbf{T_0...T_N}$ by a 4/3 power nonlinearity. The scale factors 915 are then used to scale sets of transformation coefficients $\mathbf{T_0...T_N}$ by the quantizer step size.

[0079] Additionally, tools including the mono/stereo 920, prediction 923, intensity stereo coupling 925, TNS 930, and filterbank 935 can apply further functions to the sets of transformation coefficients $\mathbf{T_0...T_N}$. The gain control 950 transforms the transformation coefficients $\mathbf{T_0...T_N}$ into the time domain signal A(t). The gain control 950 may transform the transformation coefficients $\mathbf{T_0...T_N}$ by application of the Inverse MDCT (IMDCT), inverse window function, window overlap, and window adding, for example, however other mathematical functions may be applied to the transform coefficients $\mathbf{T_0...T_N}$. The gain control 950 also looks at the flag 775. The flag 775 is a bit that may be either on or off, i.e., having binary digital value of 1 or zero, respectively. For example, if the bit is on, this indicates that the audio signal is music, and if the bit is off, this indicates that the audio signal is speech, or vice versa.

[0080] If the flag 775 indicates that the audio signal is speech the gain control may discard frequency coefficients greater than 4,000 Hz and then perform the decoding by performing the Inverse MDCT function, for example. The gain control 950 may also report results directly to the audio processing unit 999 for additional processing, playback, or storage.

[0081] Another music/speech classifier 966, such as the speech/music classifier 500 disclosed in **Figure 5**, may be provided at the decoder 900, so that in the circumstance where the signal has been received at the decoder 900 without being classified as one of speech or music, the signal may then be classified. The signal and the speech/music classification apparatus 966 output can be passed to an audio processing unit 999 for processing, playback, or further analysis, as desired.

[0082] The foregoing description of the exemplary embodiment of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not with this detailed description, but rather by the claims appended hereto.

[0083] While the invention has been described with reference to certain embodiments, it will be understood by those skilled in the art that various changes may be made and equivalents may be substituted without departing from the scope of the invention. In addition, many modifications may be made to adapt a particular situation or material to the teachings of the invention without departing from its scope. Therefore, it is intended that the invention not be limited to the particular embodiment disclosed, but that the invention will include all embodiments falling within the scope of the appended claims.